DOCUMENT RESUME

ED 102 208 95 TM 004 509

AUTHOR	Bradley, Robert H.; Caldwell, Bettye M.
TITLE	Issues and Procedures in Testing Young Children. TM Report No. 37.
INSTITUTION	ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
REPORT NO	TM-37
PUB DATE	Dec 74
CONTRACT	OEC-0-70-3797(519)
NOTE	16p.
EDRS PRICE	MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS	Affective Tests; *Children; Civil Liberties; Cognitive Tests; *Decision Making; Elementary School Students; Evaluation Criteria; Observation; Preschool Education; Psychomotor Skills; Test Bias; *Testing; Test Reliability; *Test Selection; Test Validity

ABSTRACT

Because of the developmental characteristics of young children, the potential user of tests for educational evaluation needs to be keenly alert to the kinds of decisions which can be made on the basis of testing and to the limitations of testing when young children serve as subjects. The choice of a test to be used will depend on the type of decision to be made. Several decision types are discussed: program planning and evaluation, screening, and administrative decisions. After considering the type of decision to be made as a factor in test selection, several test characteristics should be considered: practical criteria, including relevance, scope, timeliness, importance, efficiency, and credibility; validity--content, discriminant, criterion, content, and edumetric; test reliability; and test bias. Cognitive, affective, and psychomotor domains are discussed in relation to determining what kinds of test characteristics are important in assessing for a particular educational decision. Some strategies which offer effective ways to obtain more ecologically valid assessment data are presented along with comments on other testing methods, human rights, and achievement tests. (PC)

ED102208

ERIC ERIC CLEARINGHOUSE ON TESTS, MEASUREMENT, & EVALUATION
EDUCATIONAL TESTING SERVICE, PRINCETON, NEW JERSEY 08540

TM REPORT 37 DECEMBER 1974

ISSUES AND PROCEDURES IN TESTING YOUNG CHILDREN

Robert H. Bradley and Bettye M. Caldwell*

INTRODUCTION

Much ado has been made of the Age of Accountability in education. The assumption that educational programs should be accountable to their sponsors and that some objective evidence should be offered to prove that the programs actually accomplish what they claim to be able to accomplish has been one of the least popular ideas to impact professional education in recent years. There is no doubt that extravagant claims are sometimes made for unproven educational programs. For example, just a decade ago, proponents of early childhood education were suggesting that academic achievement and personality characteristics could be improved and subsequent dropout rates reduced by having children start school prior to the officially sanctioned ages of five or six. Since any new program, or any significant modification of an existing program, can cost the citizenry a great deal of money, and since the history of the field of education has been characterized by a progression of panaceas offered by one expert or another, it is not too surprising that in this era of greater scientific and administrative sophistication, clamors should be heard demanding that educators demonstrate the value of any large-scale endeavor which will consume a significant portion of public monies.

Implicit in the concept of accountability is the notion that the merits of educational programs can be demonstrated objectively and convincingly. Such demonstrations rely on formal evaluations, evaluation being defined as the process of delineating, obtaining, and providing useful information for making decisions.

In the field of education, the handmaiden of evaluation has been testing—sometimes of individuals but far more often involving groups of children. Although testing is only one of many sources of information useful for evaluation, it is the method which has been most widely employed, perhaps primarily because of the magnitude of the evaluative task when millions of individuals are involved. For this reason, it is important to make clear that testing and evaluation are not synonymous. Moreover, as the number of children to be evaluated increase, intensity of the testing effort must, of necessity, be decreased. Because of the greater verbal and literary sophistication of older children and adults, this diminution of intensity need not necessarily complicate the testing process. However, because of the developmental characteristics of young children, the potential user of tests for educational evaluation needs to be even more alert to the kinds of decisions which can be made on the basis of testing and to the limitations of testing when young children serve as subjects.

For many educators, the world of testing is confusing and frustrating. It is easy to be overwhelmed by the profusion of forms that tests take, the variety of content they contain, the proliferation of uses to which they are put, and the diversity of conditions under which they are administered. Terms such as projective tests, paper-and-pencil tests, achievement tests, diagnostic tests, personality tests, psychometric tests, edumetric tests, true-false tests, multiple-choice tests, norm-referenced tests, criterion-referenced tests, standardized tests, and teacher-make tests confront the educator frequently and produce bewilderment about the nature of the testing process. This bewilderment is easily compounded when one is concerned with the value or the limitations of testing very young children. With young children, the external criteria against which the usefulness of testing can be evaluated are either missing or less easily recognized. Furthermore, it is much more difficult to determine with young children the extent to which test results are distorted by factors other than the abilities or traits one is attempting to measure. Thus, educators planning to develop testing programs for young children need to be

*The authors' work is supported in part by Grant No. H-508 from the Office of Child Development and by a grant from the Carnegie Corporation.

This publication was prepared pursuant to a contract with the National Institute of Education, U.S. Department of Health, Education and Welfare. Contractors undertaking such projects under government sponsorship are encouraged to express freely their judgment in professional and technical matters. Points of view or opinions do not, therefore, represent official National Institute of Education position or policy.

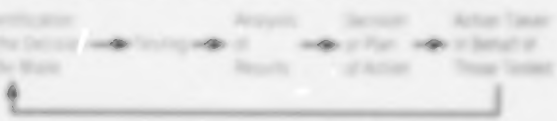
alert not only to the issues pertinent to testing with any age group but especially to the difficulties and limitations associated with testing of the very young.

All tests are *micro-experiments* (Caldwell, 1972) in that they sample skills or attitudes presumed to be representative of a larger repertoire of skills and attitudes hypothesized by the experimenter (tester) to be useful indicators of achievement or development. This is true in the case of a well organized and conceptualized collection of micro-experiments (test items) such as the Stanford-Binet Intelligence scales or in the case of a single micro-experiment test such as an attempt to determine if a child has attained the concept of conservation of mass. Tests are micro-experiments in which some capability, attitude, or behavior is examined to see if it is present or absent under a particular set of conditions. Such micro-experiments are administered for only one valid reason—the making of decisions about some course of action to be taken which in some way will involve the child who has been tested. Just as a medical test is given in order to help make a decision as to whether a certain form of medical treatment might be necessary or advisable, so are educational tests administered to help make educational decisions.

TESTING FOR DECISION-MAKING

The model below illustrates the typical manner in which tests are used as part of a decision-making process in a democratic system of education.

Model of Decision-Making in a Democratic System of Education



First, the educator identifies two particular decisions to be made. Does Angela Durkheim need special individual help due to a learning disability? Does the school district need more kindergartens? Second, a test is used to help make these decisions. In the case of the decision about Angela, a battery of tests aimed at diagnosing her specific learning disability and suggesting appropriate remediation might be administered. In the case of the decision about the additional kindergartens, each entering first grader in the district might be given a readiness test to determine the percentage of entering children identified as significantly below the expected achievement level. Third, the results of the tests are analyzed in relation to the decisions to be made. Angela has a rather severe perceptual deficit, and 40 percent of the students in the district are at least one year below entering first-grade level. Fourth, decisions are made about what to do. Angela should be given attention for her perceptual problem. New kindergartens are needed in areas of the city with a high proportion of children identified as not ready to begin first-grade work. Fifth, action is taken as a result of the decision. Angela is placed with a learning disabilities specialist who designs a special program for her perceptual problem, and the district opens four new kindergartens in schools with the highest percentage of poor achievers. Finally, after the passage of enough time to permit some results to appear, the entire cycle should begin again, with further decisions and testing. Has the remediation program helped Angela? Are the new kindergartens helping to improve scholastic readiness in entering first graders? There are new but related decisions which will require a second cycle of decision making.

These two rather oversimplified examples in no way exhaust the list of educational decisions that may fruitfully involve tests as part of the decision-making process. However, they do show the general manner in which tests should function in the making of educational decisions. They should seldom, if ever, be the only thing considered as these decisions are made, and they should not be given undue, some clear benefit for children and the community can be anticipated as an outcome.

Types of Decisions

Tests provide useful information for making many different types of educational decisions. In a given decision situation, the choice of a test to be used and the utility of the one chosen will depend on the type of decision to be made. For this reason, we shall consider in detail different types of decisions relating to the education of young children.

Program Planning and Evaluation. A large percentage of the decisions made in educational settings involve decisions relating to planning and evaluation. Some of these decisions relate to individuals ("Is Tommy Harrison making the kind of progress expected in reading?" or "Should Clarence Bell go on to the next level of instruction in math?"). Other decisions relate more directly to programs ("Should the present language arts program be changed to better meet the needs of the children participating?" or "Is the present kindergarten program accomplishing its objectives?"). Bloom, Hastings, and Madaus (1971) have classified these types of decisions as formative and summative evaluation decisions.

The first major type of decision relating to program planning and evaluation is called formative evaluation. Once a child has entered an instructional program matched to his/her individual needs and capabilities, it is important to monitor the child's progress during the program and to check for errors in the program. These two functions are called formative evaluation and can be distinguished from the functions of diagnostic evaluation, such as classifying students and determining underlying causes of difficulty. Essentially, a formative evaluation is designed to allow an educator to decide whether modification is needed in instruction to enable children to develop in a satisfactory manner.

In general, the instrument that determines whether the child possesses certain prerequisite characteristics which make him/her a good candidate for a particular program is different from the instrument which locates where in the program a child is experiencing difficulty. In many instances, diagnostic tests are concerned with individual differences which predispose an individual to receive a certain kind of treatment. Formative tests, by comparison, are primarily concerned with changes within an individual in skills, attitudes, and so on, resulting from the instructional program.

Cahen (1978) has introduced an important distinction about the function of tests which is helpful for the design and selection of formative tests. He says that a test may serve either a psychometric or an edumetric function. Psychometric tests, which have dominated the testing field for many years, focus on individual differences, whereas edumetric tests reflect individual growth from one time to another. Edumetric tests and formative evaluation are very similar concepts; when one is concerned with formative evaluation, one would be wise to choose an edumetric test. To be more specific, the test designer should choose items that are maximally sensitive to individual growth (edumetrically useful). The reason for choosing edumetrically efficient tests is that education is a purposeful enterprise, one aimed at fostering individual development. Educators should not be concerned merely with showing how individuals stand in relation to one another. Therefore, the best items for formative evaluation are not those that approximately 50 percent of the students "pass" and for which there is an approximately normal distribution. The best items are those which almost nobody passes prior to instruction and which, it is hoped, almost everybody will pass at the end of instruction (Cahen).

The concepts of edumetric testing and formative evaluation are important because program planning should often attempt to determine whether some particular type of program shows the potential for meeting educational needs of children. Thus, if one were trying to plan a program to ameliorate learning disabilities in preschool children, one would want to assess the children's progress frequently and in fairly narrow segments of activity in order to estimate whether the educational program does appear to be accomplishing its objectives. If not, then one would want to change certain aspects of the instructional program to try to get the program back on target.

The second major type of decision relating to program planning and evaluation is called summative evaluation. At the end of a program or a unit of instruction, educators are often called upon to make decisions related to the final performance of participants. Should the program be continued? What grade should an individual get? Should an individual be certified as having certain skills and knowledge? Should a participant enter the next level of the program? Did the program accomplish its objectives? What additional effects did the program have other than those intended? Such decisions are related to what Cronbach (1971) calls summative evaluation. In formative evaluation, the focus is on a test which measures improvement in all the specific skills to be achieved during the unit of instruction. In summative evaluation, the focus is more general—the educator wishes to obtain a representative sample of performance on program objectives and related variables.

A summative evaluation usually involves measuring level of mastery or attainment at the end of the total instructional period. Since individuals are likely to differ in terms of how much they have benefitted from a program, educators will often find it useful to include items which assess a fairly wide range of performance for each content area in a summative test. In this way, the educator can obtain a more accurate estimate of the program's effects on each individual. We might add that including items which have a range of difficulty is useful regardless of whether the individual's performance is compared to norms or to more specific performance criteria.

In thinking of evaluation, it might be useful to consider the taxonomy offered by Morgan (1972). In evaluation, the problem is determined by the situation. Many people may be involved in defining the evaluation, including those who participate in establishing the philosophy and objectives. Morgan also contends that the outcomes of an evaluation are not intended to be generalized to other activities. Findings apply only to the particular program evaluated. The goals and objectives of the program are used as criteria to determine the extent to which the purposes of the program have been achieved.

In the past, it has been common practice to use standardized achievement tests in summative evaluations. Tyler (1974) offers two major criticisms of this type of norm-referenced instrument when used for such a purpose. First, norm-referenced achievement tests include too small a sample of exercises appropriate for

[illegible] the learning of children who markedly deviate from the average. The tests include a low index of the level where most disadvantaged children are learning that changes in test scores due to improved capabilities cannot be distinguished from those due to chance variance in performance. Second, a teaching method or set of instructional materials is generally designed to improve learning that can or diagnose it successfully. If an [illegible] percentage of 4 or 5% percent in student learning can often be considered a success. To measure such differences in the capabilities of students would require a gain score of measurement that is rarely attained with a survey instrument such as a standardized achievement test. In a comparison, more achievement tests cannot give valid information about the differential effectiveness of teaching methods or instructional materials unless the differences are large.

Screening. In the case facilities, kindergartens, and elementary schools, considerable attention is devoted to screening for developmental difficulties. The screening decision in these settings is similar to the selection decision in industrial settings (see Cronbach [illegible]) for a discussion of selection decisions). Cartwright and Mellon [illegible] point out that the essential decision is a dichotomous one—a child either goes into a program or he does not. Therefore, that instrument is most useful which does the best job of balancing errors in deciding whether a handicapping condition is present or absent for the individual child. To be more specific, a good screening instrument is one which has both a low percentage of false positives (those identified as having the condition who, in fact, do not have it) and a low percentage of false negatives (those identified as not having a condition who, in fact, do have it). The relative importance of the two types of errors will depend on several factors including the need for specialized programming, the opportunity for [illegible] diagnosis, and the availability of programs for the condition.

Screening instruments for young children are likely to be concerned with identifying children who, if permitted to enroll in regular classrooms without some instructional adaptation geared to their particular disabilities, could be expected to experience frustration and failure rather than success. Examples are developmental delay or mental retardation, learning disabilities, or early behavioral dysfunction. The use of screening in a particular educational setting does not mean that the children are going to be identified and separated out in any organizational. With the advent of mainstreaming handicapping children with special learning needs in regular classrooms with other children, screening programs are more likely to be used to alert teachers to the special needs of some of the children and to encourage greater individualization in classroom instruction.

Diagnostic. In many educational systems, it is standard procedure to follow screening tests with more comprehensive diagnostic testing. This more careful examination of the individual leads to a decision which differs from the first decision made in screening. That is, the use of diagnostics is not to determine whether or not a person should be placed in a program but rather what specific type of instruction is the best for the child. In contrast, the purpose of screening is to determine whether a [illegible] is needed. The purpose of diagnostic testing is to determine what kinds of instruction is needed. Diagnostic testing [illegible] What could be a classic [illegible] [illegible] [illegible] is a discussion of the issues and a [illegible] decision. This distinction is made more because a test of which does an excellent job of identifying a child with a [illegible] problem may be virtually useless in prescribing [illegible]

Placement. A diagnostic test is used as a basis for placing [illegible] different programs or treatments. An [illegible] value of validity of the test lays in determination of most [illegible] For instance, because those whose reading difficulties stem from a visual perception problem will those of those reading difficulties stem from poor word attack skills, [illegible] [illegible] appropriate prescriptions for remediation of the specific problem of the [illegible]

Administrative Decisions. Wiersma [illegible] states that [illegible] decisions should [illegible] on the [illegible] of management process. Regional management is [illegible] at [illegible] an even different administrative [illegible] Wiersma goes on to say that the management decision process is quite complex and is a series of sequential [illegible] At the first levels of management, such as the local [illegible] of a decision making, are driven to extremes. Moreover, at [illegible] the needs of administrators at each level differ, so there is often a real problem of communication [illegible] between persons at the different levels. Because of this [illegible] Wiersma believes that even administrative [illegible] will have to solve a single optimization [illegible] Thus, in effect choosing a test is a means of providing information for a specific administrative decision. It is [illegible] test to others say they provide the best of [illegible] [illegible] in the particular administrative unit [illegible] of the decisions to bring [illegible] [illegible] which [illegible] [illegible] [illegible]

Like the other educational decisions, administrative ones [illegible] made with respect to a specific [illegible] Administrative decisions concern whether a new test [illegible] or replace a current one, whether [illegible] a [illegible] program, [illegible] [illegible] [illegible]

children's competencies than do typical testing procedures. One such method is the micro-experiment. Caldwell (1972) has stated that each test item from an infant development scale is itself a test. That is, it is an experiment whereby one can judge whether a child can or cannot do a particular thing. A single test item by itself can often be more ecologically valid than the same item in the context of an entire scale. The main reason for this is that a single test item is not as likely to be encumbered with irrelevant difficulty as is an entire scale of items.

Messick and Anderson discuss four sources of irrelevant difficulty which lead to bias in tests. One such source cited by the authors is test format. Suppose, for example, that a test requires the child to read each item. Reading itself may be an irrelevant difficulty. When a child is confronted by a single item, it is much easier for the teacher to overcome this kind of difficulty. A second source of irrelevant difficulty is that items are sometimes more germane to one group than to another. When items are given individually, the teacher or educator can be selective in terms of giving some items to certain groups and some items to others. Testing conditions can also make some individuals feel anxious, threatened, or alienated. When test items are given by themselves, the test will seem more like everyday problems which generally occur singly rather than in groups. Thus, such a test is less likely to cause anxiety in children. Similarly, "test wiseness," a fourth source, would appear to be less of a factor of irrelevant difficulty with single items since the problem to be solved in an individual item is more like the kind of normal problem-solving situation that every child (regardless of experience) encounters. Many such micro-experiment tests are particularly useful in the investigation of conservation skills many of which can be assessed in gamelike situations that are more like the real world than the usual classroom testing situation.

The performance test is another alternative to typical testing procedures used in educational settings. In such tests, the individual must demonstrate a competency rather than answer questions about it. Performance tests have been used for many years, particularly to examine adults' ability to do a particular job after training. They have also been used to assess adolescents' skills in school settings. The performance measure also offers a potentially useful method of assessing the capability of young children, although it has not been used often for this purpose. Performance tests have the advantage of being more ecologically valid as a means of assessment than typical testing procedures. Boyd and Shimberg (1971) discuss the advantages of using performance tests as measures of classroom evaluation. They also discuss the means through which one can develop a performance test. To quote them, "Most of us recognize that there is a fundamental difference between knowing about a job and being able to do the job. Knowledge of a job is really an essential ingredient for doing a complex job correctly, but while it is a necessary condition, it is rarely a sufficient condition for doing a satisfactory performance. . . A person may be able to bluff on a written test, but he can seldom carry off a successful deception when a realistic performance test is required. One of the great virtues of the performance test is its impressive face validity and credibility, because the task one must do so closely resembles the job itself" (p. 3).

A performance test, conducted in an actual or simulated real-world setting, can be particularly useful for assessing certain kinds of capabilities in children. If one is interested in examining communication skills in children, for example, one might better assess this by placing them in a situation where they have to demonstrate their communication skills than by testing their usage of grammar in a written examination. The same is true with mathematics. Setting up situations in which one has to compare prices on food, compare percentages of nutritional ingredients in food, make correct change in monetary transactions, judge whether one is being hoodwinked in monetary transactions, and so forth, may be an excellent means of assessing certain mathematical capabilities in young children. Similarly, the kind of math proficiencies demonstrated in the playing of certain children's games may also be excellent indices of mathematical ability. Boyd and Shimberg (1971) state that "Whether one starts with a job analysis or with behavioral objectives which are originally derived from such an analysis, one must decide which elements are crucial to success. It is from these critical elements that one should select the task to be used as a measure of performance. Because performance is generally a slow and time-consuming process, only a few of the critical elements can be included. One must decide which ones are really crucial" (p. 5).

GENERAL CONSIDERATIONS

Human Rights

In considering the proper function of testing in education, one quickly realizes how often tests have been used improperly. One of the most lucid discussions of the kinds of abuses which occur when tests are not used properly appears in a recent article by Mercer (1974). She lists five rights of children that are frequently violated by present assessment and educational practices. In brief, those rights are as follows:

(1) *The Right To Be Assessed as a Multi-Dimensional Human Being.* Mercer found that many children are classified as retarded on the basis of IQ tests alone, even though IQ tests assess only a fraction of the total number of human skills and capabilities. For some children, retardation is school-specific. Their competencies in other situations are quite adequate.

(2) *The Right To Be Fully Educated.* For many children, educational assessment is a prelude to being stuck in a track or class which in no way meets their individual educational needs.

(3) *The Right To Be Free of Stigmatizing Labels.* The problems associated with being labeled as retarded, speech-impaired, learning-disabled, and so forth are clearly documented. Mercer argues that the learning environment must be carefully structured so as to avoid the negative stereotyping which accompanies such labels.

(4) *The Right To Ethnic Identity and Respect.* As a general rule, standardized intelligence tests and achievement tests assess those competencies valued in Anglocentric societies while ignoring many of the competencies valued in other cultures. Such tests often provide "a mechanism for blaming children and their families when the educational program of the school fails" (Mercer, 1974, p. 137).

(5) *The Right To Be Evaluated within a Culturally Appropriate Normative Framework.* Many assessment procedures employ a single normative framework for interpreting the scores of all children. Those tests ignore the fact that the experiences of many minority group children differ widely from those of the dominant cultural group. Thus, for minority group students, comparison to the norm can be misleading.

To this list we would add at least one additional right that is especially applicable to young children—the right not to be permanently labeled on the basis of testing done during the preschool period. Present assessment techniques may be even more hazardous with very young children than they are with slightly older children and adults. Problems of reliability and validity (all types) are much greater with the very young child. If a child is growing and changing, do we want a test that produces identical results across time, one that has high test-retest reliability? As fatigue builds up quickly and militates against a long testing procedure, should we expect high split-half reliability? As large representative groups of young children have seldom been assembled, as can be the case with public school samples of older children, norm-referenced tests seldom have norms based on statistically acceptable samples. Criterion validity is difficult to determine, as widely accepted outside criteria can seldom be identified. For example, suppose one wanted to develop a test to measure emotional disturbance in preschool children. Presumably one might use psychiatric diagnosis as an acceptable outside criterion against which to validate the test. But most psychiatrists would feel doubtful about identifying those children whose behavior might justifiably be labeled emotionally disturbed because immaturity is an outstanding characteristic of such behavior, and all very young children are, by definition, immature.

These hazards do not mean that one has no right to apply tests to young children; they do suggest, however, that the person who uses such test data for educational decision making must anticipate quick obsolescence for any test data obtained on young children. That is, a cardinal rule for the testing of young children should be frequent retesting, in that fairly large changes in behavior can be expected in short periods of time. No child should be labeled permanently on the basis of test data obtained during the early childhood period.

Many educators have voiced concern over the abuses made of standardized testing by school personnel. Their concern points up a more general issue regarding the use of any kind of information in educational settings. As educators we must ask ourselves: What am I obliged to do in terms of gathering and disseminating school-relevant information? In a democratic system of education, the one who orders that information be obtained is the agent chiefly responsible for the information. Take the case of a school administrator. If an administrator wants a test to be given to a group of students by their teacher, he/she must make provisions for informing the teachers of the correct ways of obtaining and interpreting the information from the test. The information must then be used exclusively for the benefit of the child and the community. One interesting recent legal ruling aimed at improving the use of information in educational institutions is the Family Educational Rights and Privacy Act of 1974. This law guarantees the parents' right of access to their children's school records. The courts have deter-

mined that parents not only have the right to know what is in the record but also the right to challenge the veracity and appropriateness of the information it contains. This landmark legislation may prevent a repetition of many of the abuses associated with testing and other school-related information.

Achievement Tests

Mercer's insistence that educators test in such a manner that the rights of children are protected offers one more strong argument in favor of deemphasizing the use of standardized, norm-referenced tests when making educational decisions. As discussed previously, Tyler contends that standardized achievement measures are seldom adequate as summative tests, especially when disadvantaged children are involved; and Carver convincingly demonstrates the weaknesses of many norm-referenced measures when used to assess improvement in capabilities. This is not to say that standardized measures should be abandoned. On the contrary, they are often a wise choice because of their meticulous construction and convenience. What is important is that educators be judicious in employing such instruments, with an eye toward their appropriateness for making a particular decision.

One additional suggestion we might offer to educators is that they not rely too heavily on a single test score when making decisions. A battery of tests often provides a broader, more precise base of information. Unfortunately, although much lip service has been paid to the wisdom of using multiple measuring instruments, little training has been given on how to use them and even less data collected on their actual effectiveness.

POSTSCRIPT

At the beginning of this paper, we indicated that accountability has become a watchword in education. Educators are more mindful of the need to demonstrate how educational programs are making good their claims. Testing, probably because of its long and rather successful history in education, has been the chief evidence-gathering vehicle for establishing accountability. This heavy dependence on the use of test information is currently being questioned both by those within the field of education and by those without. Complaints are being made that tests may be inadequate as a basis for making many educational decisions. It is perhaps fair to say that testing has entered its own Age of Accountability—educators can no longer be arbitrary when using tests. The person tested or his family may offer a rebuttal to what the tests have found. In the still largely mysterious world of young children, it is even more essential that educators heed these warnings.

REFERENCES

Allen, K.E., Rieke, J., Dmitriev, V., & Hayden, Alice H. Early warning: observation as a tool for recognizing potential handicaps in young children. *Educational Horizons*, 1972, Vol. 50, No. 2, 43-54.

Anderson, Scarvia B., Ball, S., Murphy, Richard T. *Encyclopedia of educational evaluation*. San Francisco: Jossey-Bass, 1975.

Bloom, Benjamin S. (Ed.). *Taxonomy of educational objectives handbook I: cognitive domain*. New York: David McCay, 1956.

Bloom, Benjamin S., Hastings, J. T., & Madaus, George F. *Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill, 1971.

Boyd, Joseph L., Jr., & Shimberg, B. Developing performance tests for classroom evaluation. *ERIC/TM Report #4*, 1971.

Bracht, Glenn H., & Glass, Gene V. The external validity of experiments. *American Educational Research Journal*, 1968, 5, 437-474.

Caldwell, Bettye M. Practical developmental testing. *Pediatric Psychiatry*, 1972.

Campbell, Donald T., & Stanley, Julian C. Experimental and quasi-experimental designs for research on teaching. In N.L. Gage (Ed.), *Handbook of Research on Teaching*. Chicago: Rand-McNally, 1963.

Carver, Ronald P. Two dimensions of tests: psychometric and edumetric. *American Psychologist*, 1974, Vol. 29, No. 7, 512-518.

Cronbach, Lee J. *Essentials of psychological testing*. New York: Harper and Row, 1969.

Ebel, Robert L. *Essentials of educational measurement*. Englewood Cliffs, N. J.: Prentice-Hall, 1972.

Gallagher, James J., & Bradley, Robert H. Early identification of developmental difficulties. *Yearbook of the National Society for the Study of Education*, 1972, Part II.

Gronlund, Norman E. *Preparing criterion-referenced tests for classroom instruction*. New York: Macmillan, 1973.

Harrow, Anita J. *A taxonomy of the psychomotor domain*. New York: David McCay, 1972.

Herman, John L. Educational management information systems for the seventies. *Educational Administration Quarterly*, 1974, 10, 60-71.

Herzig, Margaret E., Birch, Herbert G., Thomas, A., & Mendez, Olga A. Class and ethnic differences in the responsiveness of preschool children to cognitive demands. *Monographs of the Society for Research in Child Development*, Vol. 33, No. 1, 1968.

Hieronymus, A.N. Today's testing: what do we know how to do? *Proceedings of the 1971 Invitational Conference on Testing Problems*. Princeton, N. J.: Educational Testing Service, 1971, Pp. 57-68.

Krathwohl, David R., Bloom, Benjamin S. and Masia, Bertram B. (Eds.) *Taxonomy of educational objectives handbook II: affective domain*. New York: David McCay, 1964.

Mercer, Jane R. A policy statement on assessment procedures and the rights of children. *Harvard Educational Review*, 1974, 44, 125-141.

Messick, S., & Anderson, Scarvia B. Educational testing, individual development, and social responsibility. *The Counseling Psychologist*, 1970, 2, 80-88.

Morgan, Don L. Evaluation: a semantic dilemma. *Educational Technology*, 1971, Vol. 11, No. 12, 46-48.

Snow, Richard E. Representative and quasi-representative designs for research on teaching. *Review of Educational Research*, 1974, 44, 265-292.

Stufflebeam, Daniel L., Foley, Walter J., Gephart, William J., Guba, Egon G., Hammond, Robert L., Merriman, Howard O., & Provus, Malcolm M. *Education evaluation and decision making*. Itasca, Ill.: Peacock, 1971.

Tyler, Ralph W., & Wolf, Richard M. (Eds.) *Crucial issues in testing*. Berkeley, Calif.: McCutchan Publishing Corporation, 1974.

Tyler, Ralph W. The use of tests in measuring the effectiveness of educational programs, methods, and instructional materials. In Ralph W. Tyler and Richard M. Wolf (Eds.), *Crucial Issues in Testing*. Berkeley, Calif.: McCutchan Publishing Corporation, 1974.